ISKO. CAPÍTULO ESPAÑOL. CONGRESO (7º: 2005: BARCELONA) 534-548

# Interoperabilidad semántica de ontologías basada en técnicas de procesamiento del lenguaje natural

M. Teresa Romá-Ferri,[1] mtr.ferri@ua.es
Manuel Palomar,[2] mpalomar@dlsi.ua.es

1. Universidad de Alicante (Alicante, España). Departamento de Enfermería
2. Universidad de Alicante (Alicante, España). Departamento de Lenguajes y Sistemas Informáticos

**Resumen**
Se presenta un estudio y propuesta de interoperabilidad semántica entre ontologías del dominio de la salud basada en técnicas de procesamiento del lenguaje natural. El objetivo fundamental ha sido el desarrollo de un algoritmo de interconexión semántica entre los términos de dos ontologías solapadas y heterogéneas, denominadas «fuente» (*Clasificación internacional de enfermedades*, 9ª revisión, modificación clínica: CIE-9-MC) y «diana» (esquema jerárquico de la asignatura *Enfermería Materno-Infantil*: EMI). Esta propuesta permite emparejar semánticamente ontologías, a partir de la reutilización de otro recurso ontológico (*WordNet español*), sin destruir o modificar la semántica de identidad de cada una de las ontologías involucradas. El modelo presentado puede permitir al usuario acceder a la información que necesita en otra clasificación jerárquica, sin precisar de un entrenamiento referido a la conceptualización de cada sistema, pues utilizaría la ontología «diana» con la que está familiarizado para su aplicación a la recuperación de información.

**Palabras clave**
interoperabilidad semántica, procesamiento del lenguaje natural, ontologías, algoritmos de negociación de significados, interconexión semántica, análisis conceptual, información en ciencias de la salud, enfermeros como usuarios de la información

## 1 Introducción

En las últimas décadas, el conocimiento ha sido reconocido como uno de los recursos fundamentales de las organizaciones, siendo valorado como la materia prima esencial. Como materia prima tiene todo el potencial, pero precisa ser transformada para ser funcional. Es decir, el conocimiento tácito residente en las personas que componen una organización debe pasar a ser un conocimiento explícito para ser compartido por todos (Nonaka 1994). Una de las formas de compartirlo es por medio de la creación de documentos, repositorios, bases de datos, etc. que son almace-

nados en ordenadores. En el proceso de almacenamiento, los documentos, por ejemplo, son clasificados para su posterior recuperación. Estos sistemas de clasificación también pueden ser entendidos con conocimiento explícito. Consecuentemente, la clasificación de documentos electrónicos es una de las actividades que va ganando protagonismo en toda organización. Esta actividad, para que sea eficaz y eficiente, precisa de sistemas de clasificación que den respuesta tanto a los objetivos como a las expectativas de la propia organización. Aunque, al mismo tiempo, precisa de sistemas que faciliten compartir la información corporativa contenida en su clasificación jerárquica, con la información que procede de otras clasificaciones jerárquicas, lo que impone la necesidad de un intercambio fluido entre los proveedores y los usuarios/clientes y, en este caso, a través de ordenadores (ITAMI 1987).

Sin embargo, el mismo conocimiento puede ser representado en cada organización, e incluso por cada individuo, desde distinto puntos de vista, ya que sus necesidades no tienen que ser coincidentes. La necesidad de información se suele reflejar en algún tipo de clasificación que permite operar con el conocimiento local o interno de la organización, normalmente pensado para ser utilizado por las personas. Un recurso que facilita la representación del conocimiento a nivel computacional son las ontologías (GRUBER 1993, LOZANO 2002). En este sentido, las clasificaciones jerárquicas contienen el vocabulario de representación; es decir, son la ontología del dominio o tema de interés. Sin embargo, no es el vocabulario en sí el que se califica como ontología sino las conceptualizaciones que los términos estructurados en la jerarquía describen (CHANDRASEKARAN 1999, FELIU 2002, SWARTOUT 1999). Las ontologías han sido creadas con propósitos diversos y relacionados, muy frecuentemente, con tareas específicas como la clasificación de documentos. No obstante, el inconveniente que surge es que estos sistemas de información, aún perteneciendo al mismo dominio, como el de la salud (DE LA CUEVA 2001, RICHART 2000), son heterogéneos respecto a la conceptualización o la semántica que incluyen. Tan sólo rescindiéndonos al ámbito de la salud (aunque esta situación también es generalizable a otros ámbitos) las diferencias entre los sistemas de información son manifiestas tanto desde el punto de vista de su estructura (esquema o clasificación jerárquica) como de su cobertura (contenido y finalidad de éste) o de la granularidad (detalle y especificidad de sus categorías conceptuales) (FELIU 2002, GANGEMI 1998, LOZANO 2002, RODRÍGUEZ 2000, ROMÁ-FERRI 2004), lo cual obstaculiza el proceso o el uso de aplicaciones cuyo propósito es compartir la información que sustenta cada uno de los sistemas.

No obstante, si aceptamos, desde el punto de vista de la gestión de conocimiento (en concreto, respecto a sus objetivos de actualización, generación y difusión de nuevo conocimiento), que lo importante para intercambiar o compartir la información es la interpretación de los esquemas de clasificación que forman parte de los sistemas de información, en lugar de buscar una mayor uniformidad conceptual deberíamos centrarnos en el problema de la interoperabilidad semántica (BOUQUET 2002). Es

decir, deberíamos valorar la posibilidad de interconexión (*mapping*) entre las diferentes conceptualizaciones que se representan en las clasificaciones jerárquicas. Para resolver este problema práctico se ha optado por diversas soluciones, y así han sido recogidas en la literatura científica, evidenciando las ventajas y desventajas de cada una de ellas en función de la propia jerarquía de clasificación y de las necesidades de información organizacional (*semantic coordination, mapping between domain models, semantic mediation, ontology merging, ontology integration, ontology alignment, integration of hierarchical categorization*) (BOUQUET 2002, 2003, MAGNINI 2003). Aunque la metodología de negociación de significados (*meaning negotiation)* puede ser una opción, ante la necesidad de interconexión de clasificaciones jerárquicas solapadas y heterogéneas, ya que su proceso habilita a los agentes a descubrir las relaciones existentes entre ambas, sin destruir o modificar la «semántica de identidad» de las ontologías involucradas (BOUQUET 2002, 2003).

El trabajo que presentamos tiene como objetivo principal proponer un algoritmo de interconexión semántica entre dos ontologías del dominio de la salud, la ontología local fuente (*Clasificación Internacional de Enfermedades*, 9ª revisión, Modificación Clínica: CIE-9-MC[1]) y la ontología local diana (esquema jerárquico de la asignatura *Enfermería materno-infantil*: EMI), utilizando técnicas de procesamiento del lenguaje natural (PLN). Para ello establecemos los siguientes objetivos específicos: i) definir un método de análisis del conocimiento oculto en la jerarquía conceptual y hacerlo explícito para proporcionar una correcta interpretación de sus conceptos; ii) definir un algoritmo que retorna una interpretación de cada nodo de una jerarquía en términos de una fórmula lógica; y, iii) definir un algoritmo automático de negociación de significados que permita la interoperabilidad entre ontologías solapadas y heterogéneas del ámbito de la salud.

## 2 Propuesta de interconexión semántica

### 2.1 Método

Las ontologías locales son entendidas como contextos y son representaciones parciales y aproximadas del mundo desde el punto de vista de un individuo o grupo (BOUQUET 2002 y 2003). En este trabajo consideramos como tales la clasificación CIE-9-MC y el esquema jerárquico del programa de la asignatura *Enfermería Materno-Infantil* (EMI). Ambas conceptualizan una parte común del mundo, en concreto del ámbito de la salud, pero la representación que se incluye del parto no está representada de igual forma, aunque existen superposiciones o coincidencias entre ambas. La jerarquía de conceptos (JC) es el modelo de referenciación general que se utilizará. Las JC son construidas desde un conjunto de etiquetas (E). A su vez, en dicho conjunto general diferenciamos dos sub-

3. En lengua inglesa se le conoce por las siglas ICD-9-CM.

conjuntos disjuntos: el de las «etiquetas de los conceptos» (Ec) y el de las «etiquetas de las relaciones» (Er). Las etiquetas son expresiones del lenguaje natural del ámbito de la salud.

Podemos definir una «JC» como un gráfico constituido originalmente por un conjunto finito de nodos (N) y un conjunto finito de arcos directos entre los nodos (A). Tanto los nodos como los arcos tienen una etiqueta perteneciente al conjunto (E), de manera que los arcos con etiquetas de la jerarquía forman un árbol.

Los conceptos representados en la ontología local por medio de las Ec son los nodos de la JC. Las etiquetas de los nodos de la ontología local se corresponden con la rúbrica que identifica el nombre de un diagnóstico en el CIE-9-MC («embarazo múltiple», «embarazo gemelar con pérdida fetal» y «retención de un feto con complicaciones en el anteparto») y con la denominación de las unidades temáticas, de los temas y de los descriptores de los contenidos que conforman el programa de la asignatura EMI («perdida de salud y/o necesidad de hospitalización», «parto», «instrumental», «manual») (ver Figura 1).

El conjunto de Er se conforma por las etiquetas «*es un/una*», «*es parte de*» o «*es una instancia*». Donde «es un/una» representa una relación de subclase («manual *es una* forma de valoración», «distócico *es un* tipo de denominación según su evolución», «según su evolución *es un* tipo de parto», «a término *es un* tipo de denominación según el momento de producirse», «según el momento de producirse *es un* tipo de parto»). La etiqueta «parte de» representa la relación de los elementos que componen un todo, bien sea un humano, un objeto o un proceso («contracciones *es parte de* los elementos del parto», «canal del parto *es una parte de* los elementos del parto», «los elementos del parto *son parte del* parto»). Con la etiqueta «es una instancia», se representa el hecho de que un cierto individuo u objeto es una formalización de un concepto («tacógrafo interno *es una instancia* de instrumental», «tacógrafo externo *es una instancia* de instrumental»).

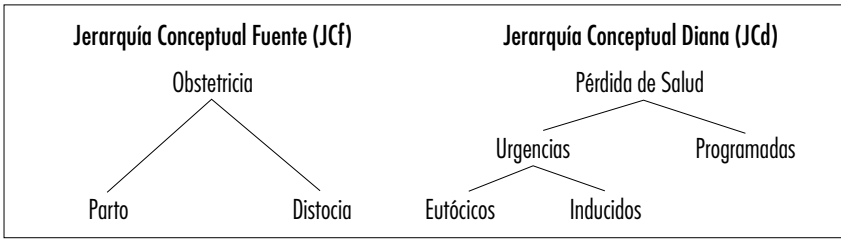



**FIGURA 1.** Representación parcial de las jerarquías de conceptos locales (ejemplo hipotético). En ellas no se incluyen las etiquetas de los arcos de conexión entre los nodos

## 2.2 Algoritmo propuesto

El algoritmo de interconexión de conceptos pretende que dada una jerarquía de conceptos fuente, JC*f*, y una jerarquía de conceptos diana, JC*d*, extraer las relaciones semánticas explícitas e implícitas de cada par de conceptos $K_f$ de JC*f* y $K_d$ de JC*d*. Por consiguiente, se debe identificar la

interconexión (*mapping*) existente entre las dos JC, entendiendo por interconexión la relación entre la «JC fuente» (CIE-9-MC) y «JC diana» (esquema jerárquico de EMI). La relación entre ambas JC será de tipo direccional; es decir, se representará la situación donde la «JC diana» importa información de la «JC fuente». La interconexión entre ambas JC no puede limitarse tan sólo a la relación entre conceptos equivalentes (relación semántica explícita). La interconexión, a su vez, debe permitir la representación entre conceptos de niveles de abstracción diferentes entre ambas JC (relación semántica implícita). Debe permitir que si un concepto «K» en una de las JC es más general («presentación podálica/nalgas») o más específico («presentación nalgas sin versión con o sin patología anteparto o parto») que en la otra JC se produzca su emparejamiento.

La propuesta de nuestro algoritmo de interconexión de conceptos en el dominio de salud se basa en los trabajos de Bouquet (2002 y 2003) y Magnini (2002a, 2002b y 2003) y en él se diferencian dos fases. La primera fase se centra en la desambiguación de las etiquetas de la «JC fuente» y «JC diana» para obtener su normalización lingüística y la segunda fase se centra en la propia interconexión entre las JC.

No obstante, en la primera fase, el análisis lingüístico sólo permite evidenciar la información implícita que deriva de la propia etiqueta del nodo, lo cual no es suficiente. Se precisa un segundo paso para interpretar la información que deriva de su posición en la jerarquía y de la relación estructural con los otros nodos del esquema de clasificación. El sentido de las etiquetas debe ser filtrado o enriquecido respecto al contexto donde se sitúa teniendo en cuenta la posición del nodo.

Por lo tanto, el algoritmo debe:

1. Dada una JC*f*
   a. Dada una etiqueta de la JC*f*
      i. Obtener su interpretación lingüística, I*f*, por medio de técnicas de PLN.
      — Análisis lingüístico.
      — Interpretación semántica.
      • Filtrado de los sentidos.
      • Composición de los sentidos.

2. Dada una JC*d*
   a. Dada una etiqueta de la JC*d*
      i. Obtener su interpretación lingüística, I*d*, por medio de técnicas de PLN.
      — Análisis lingüístico.
      — Interpretación semántica.
      • Filtrado de los sentidos.
      • Composición de los sentidos.

3. Dadas las interpretaciones I*f* y I*d*.
   a. Obtener la relación del emparejamiento por medio de la matriz I*f* x I*d*. El resultado es un subconjunto de las posibles relaciones entre

los conceptos emparejados ($K_f$ y $K_d$) que puede incluir una o varias relaciones:

- — Cuando $K_f$ es más general que $K_d$ («parto» es más general que «parto en un caso totalmente normal»)
- — Cuando $K_f$ es menos general que $K_d$.
- — Cuando $K_f$ es más general que $K_d$ y $K_f$ es menos general que $K_d$. («parto en un caso totalmente normal» es equivalente a «eutócico»).
- — Cuando $K_f$ es disjunto de $K_d$ («canal de parto» es disjunto de «objeto de parto»).
- — Cuando $K_f$ es compatible con $K_d$ («parto en un caso totalmente normal» es compatible con «postérmino» ya que el tiempo de gestión prologado es posible con una evolución de parto normal).

## 3 Análisis cualitativo del algoritmo de interconexión semántica

El propósito de la primera fase es lograr la interpretación semántica de las etiquetas de cada concepto. Las etiquetas de los nodos se corresponden con expresiones en lenguaje natural que contienen una terminología específica perteneciente al dominio de salud. Las expresiones lingüísticas son muy variadas:

- — Etiquetas simples: «distocia», «contracciones», «Braxton Hicks», «Bartholin», «PEG», «simple», «múltiple».
- — Sintagmas nominales: «parto largo», «cordón umbilical corto», «hematoma pelviano», «hemorragia postparto».
- — Sintagmas preposicionales: «retención placenta sin hemorragia», «según el número de fetos».
- — Etiquetas complejas: «enfermería materno-infantil», «trauma de perineo y vulva durante alumbramiento», «retención placenta sin hemorragia-cuidados NEOM», «anomalía fetal comprobada/sospechada que afecta a la madre».

1. El primer paso se basa en la realización del análisis morfosintáctico de las etiquetas de los nodos. Las herramientas y recursos que se utilizan son: i) Tokenizador para identificar cada token que compone la etiqueta; ii) Lematizador para realizar el análisis morfológico de cada token, lo que permitirá identificar la forma canónica, la categoría gramatical y la flexión o derivación que la produce; iii) Postagger para seleccionar, entre las categorías propuestas por el lematizador, la que le es adecuada para el lema; iv) Reconocedor de entidades para la identificación de «nombres propios» (personas, organizaciones, medida) que pueden formar parte de la etiquetas (ver ejemplos en las etiquetas simples); v) Reconocedor de siglas para intercambiar las siglas y las abreviaturas por su expresión completa (ver ejemplos en las etiquetas simples y etiquetas complejas); vi) Analizador sintáctico para la obtención de la es-

tructura de representación que describe las relaciones internas que se establecen entre los elementos que constituyen las etiquetas.

El resultado del análisis lingüístico desarrollado hasta aquí conlleva tener una representación que codificará la información lingüística de las etiquetas de cada una de las JC, tanto morfológica como sintáctica (ver Figura 2).

| «Pérdida de Salud» | | | |
|---|---|---|---|
| **ID** | 0 | 1 | 2 |
| **Token** | Pérdida | de | Salud |
| **Lema** | pérdida | de | salud |
| **Postagger** | sustantivo | preposición | sustantivo |
| **Función** | núcleo | preposición | núcleo |

**FIGURA 2.** Resultado del análisis morfosintáctico de la etiqueta del nodo «Pérdida de Salud» (*JCd*)

2. El segundo paso es el análisis semántico. Para la interpretación semántica de las etiquetas se utilizará como recurso la base de conocimiento léxica *WordNet* para el español. No obstante, como nuestro trabajo se concreta en el dominio de la salud, también se precisará del recurso *WordNet Domains* que contiene una serie de etiquetas o dominios asociados a cada synset de *WordNet 1.6* denominadas «*Subject Field Codes*» (SFC). Es decir, identifica el conjunto de palabras pertenecientes a un dominio específico, lo cual extiende la cobertura de las etiquetas de dominio dentro de la base léxica *WordNet 1.6* (MAGNINI 2000). Además, se utilizará un nuevo recurso léxico denominado «Dominios Relevantes». Este recurso ha sido obtenido a partir de la información contenida en la glosa de *WordNet Domains* y su finalidad es establecer para cada una de las palabras de *WordNet* el conjunto de dominios más relevantes. En concreto, en este último recurso hemos identificado la existencia de dominios relevantes en el ámbito de la salud y, directamente vinculados con la finalidad de nuestro trabajo, están los dominios de «medicina», «cirugía», «radiología», «anatomía», «fisiología» y «farmacología», así como los de «sexualidad», «psicoanálisis» y «biología» indirectamente (VÁZQUEZ 2004). Los dominios relevantes serán un recurso esencial para la interpretación semántica de las etiquetas de los nodos de las JC.

   El objetivo de este paso es hacer explícita toda la información semántica implícita en las etiquetas y en la misma estructura de la jerarquía. El procedimiento que se establecerá por el algoritmo es:

   a. El procedimiento toma todos los sentidos asociados a un lema, los cuales serán seleccionados y se transferirán a la tabla que contiene la estructura lingüística de cada etiqueta. La finalidad de este paso es hacer posible la asociación de cada etiqueta a una fórmula lógica que codifica la información que es externa al propio contexto que contie-

ne la jerarquía. Es decir, intuitivamente, es una aproximación de la interpretación humana de los conceptos que incluye un término.

Tomando como ejemplo el concepto «pérdida de salud», podemos encontrar que el lema «pérdida» pueda tener cuatro sentidos significativos asociados, *pérdida#1* (dejar de tener o no hallar aquello que poseía, sea por culpa o descuido del poseedor, sea por contingencia o desgracia), *pérdida#2* (desperdiciar, disipar o malgastar algo), *pérdida#3* (dejar salir poco a poco el contenido de un recipiente), *pérdida#4* (empeorar de aspecto o de salud). De igual forma, para el lema «salud» podemos localizar *salud#1* (estado en que el ser orgánico ejerce normalmente todas sus funciones), *salud#2* (condiciones físicas en que se encuentra un organismo en un momento determinado), *salud#3* (estado de gracia espiritual).

En este paso, se asocia a cada lema una fórmula lógica representando la interpretación de ese lema. Para la descripción lógica se adoptan los siguientes símbolos: $\cup$, $\cap$ y $\neg$ (conjunción, disyunción y negación) y cuyos conceptos primitivos son los synsets de WordNet asociados a cada lema de la etiqueta (ver Figura 3).

| «Pérdida de Salud» | | | |
|---|---|---|---|
| **ID** | 0 | 1 | 2 |
| **Token** | Pérdida | de | salud |
| **Lema** | pérdida | de | salud |
| **Postagger** | sustantivo | preposición | sustantivo |
| **Función** | núcleo | preposición | núcleo |
| **W-sentidos** | pérdida#1<br>pérdida#2<br>pérdida#3<br>pérdida#4 | | salud#1<br>salud#2<br>salud#3 |
| **Interpretación** | pérdida#1 $\cup$<br>pérdida#2 $\cup$<br>pérdida#3 $\cup$<br>pérdida#4 | | salud#1 $\cup$<br>salud#2 $\cup$<br>salud#3 |

**FIGURA 3.** Resultado del análisis morfosintáctico y semántico del concepto «pérdida de salud»

b. El objetivo de este paso es eliminar o enriquecer los synsets asociados a las etiquetas, pero usando el contexto en el cual se sitúa y, por consiguiente, se le puede denominar como «interpretación de contexto». Para ello, precisamos introducir la idea de «foco de un concepto». El «foco de un concepto» es el subconjunto de la jerarquía que se precisa considerar para determinar el significado de un concepto. El «foco de un concepto» depende de la estructura explícita de la representación de la JC, lo que determina los antecesores y los descendientes de un concepto.

La contextualización de la interpretación de un concepto de una JC es una formula denominada «interpretación contextual», la cual se computa combinando las interpretaciones lingüísticas asociadas a cada concepto de la JC respecto al «foco del concepto». Las dos operaciones que se precisan realizar son la de filtrado de los sentidos (contextualización vertical) y la de composición de los sentidos (contextualización horizontal). Las reglas que se utilizarán para el enriquecimiento de la interpretación o bien para la eliminación de uno o varios de los sentidos de una lista de sentidos asociados será:

— Si para un determinado sentido de un «Concepto inicial» hay un «Concepto antecesor» y el sentido del concepto antecesor es hiperónimo del sentido del concepto inicial, se mantendrá. El sentido inicial es eliminado si no hay un concepto antecesor y no hay un sentido hiperónimo.
— Si para un determinado sentido de un «Concepto inicial» hay un «Concepto descendiente» y el sentido del concepto descendiente es hipónimo del sentido del concepto inicial, se mantendrá. El sentido inicial es eliminado si no hay un concepto descendiente y no hay un sentido hipónimo.
— Si hay un «Concepto padre» del «Concepto inicial» en donde todos los sentidos del concepto padre son opuestos o complementarios al sentido inicial (frío-caliente, relajado-contraído), el sentido inicial es eliminado.
— Si el sentido del concepto del nodo del primer hermano es un merónimo (parte de, miembro de, porción/trozo/pedazo) del nodo del segundo hermano y el sentido del segundo hermano es un halónimo (hecho de, compuesto de, que contiene, que se compone) del primero, se remplazará el sentido del segundo hermano excluyendo el significado del nodo del primer hermano. Ilustramos está situación con un ejemplo existente: la «JC*f*» (ver Figura 4).

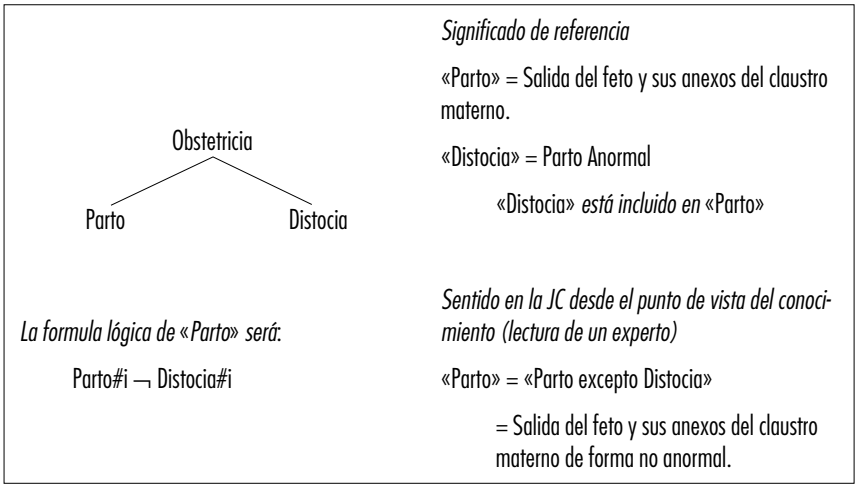



**FIGURA 4.** Composición de sentido para el concepto «parto»

La segunda fase se centra en la propia interconexión (*mapping*) entre las jerarquías conceptuales. La interconexión debe entenderse como la evidencia de la relación entre un nodo ($k_1$) en una JC*f* y un nodo ($k_2$) en una JC*d*, lo cual se reduce a un problema de verificación del conjunto de relaciones lógicas entre la fórmula de «interpretación contextualizadas del nodo ($k_1$)» y la fórmula de «interpretación contextualizadas del nodo ($k_2$)». La idea es ver los sentidos de *WordNet* que están contenidos en la formula de la «interpretación contextualizada del nodo» como un conjunto de documentos; en otras palabras, el conjunto de documentos asociados al nodo de la clasificación tratará del sentido de los conceptos contenidos en la fórmula de interpretación. Para llevar a cabo esta fase, es preciso realizar los siguientes pasos:

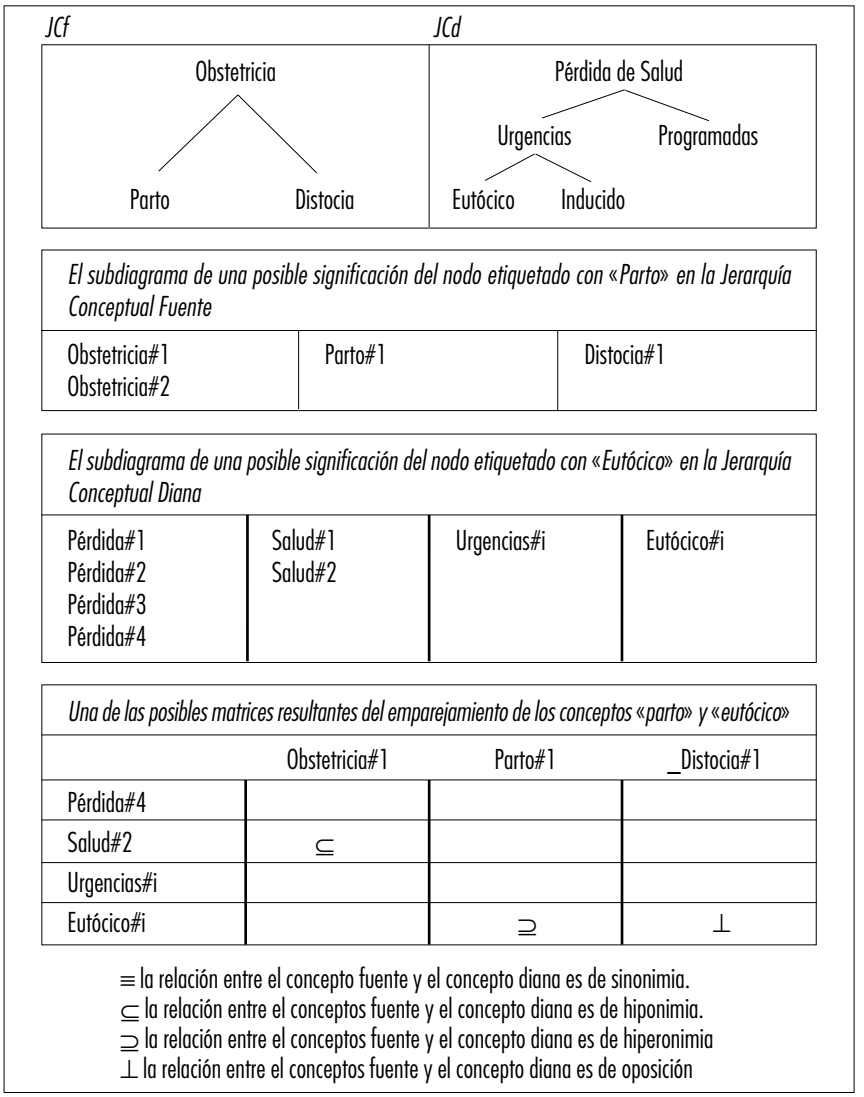
1. Elaborar la fórmula de significación contextual de cada nodo en su JC. Su contenido será una lista ordenada y refinada de la interpretación de contexto de las etiquetas de cada nodo incluido en cada rama (de acuerdo al foco del concepto). En las etiquetas de los nodos se incluye cada significado de contexto asociado a los lemas, salvo las palabras vacías (artículo, preposición, etc.). Respecto al orden, en primer lugar se incluirá la raíz de la trayectoria, seguida por sus hijos y los hijos de sus hijos hasta la etiqueta del nodo; esta representación se entiende como «el foco del concepto» (un subdiagrama del conjunto total de la JC). En el caso del nodo «Parto» de la JC*f*, una de sus fórmulas de significación contextual es (Obstetricia#1$\cap$(Parto#1$\cup\neg$Distocia#1)), mientras que para el nodo «Eutócico» de la JC*d* es ((Perdida#4$\subseteq$Salud#2)$\cap$Urgencias$\cap$Eutócico).

2. Completar la matriz de emparejamiento en cuyas filas se incluirán los componentes de la significación contextual de $k_1$ (el subdiagrama «JC fuente») y en las columnas los componentes de la significación contextual de $k_2$ (el subdiagrama «JC diana») (ver Figura 5).

   Los valores de las celdas de la matriz se obtendrán de la aplicación de las siguientes reglas, al relacionar cualquier par de sentidos de los conceptos fuente y diana ($k_1$ y $k_2$) que se pretenda emparejar:

   a. Si un sentido de $k_1$ es sinónimo de un sentido $k_2$ en WordNet, quitamos todos los sentidos diferentes tanto del concepto $k_1$ como del concepto $k_2$ y en la celda de cruce se indica su relación de equivalencia.
   b. Si un sentido de $k_1$ es hipónimo o merónimo de un sentido $k_2$ en WordNet, quitamos todos los sentidos diferentes tanto del concepto $k_1$ como del concepto $k_2$ y en la celda de cruce se indica dicha relación.
   c. Si un sentido de $k_1$ es hiperónimo o halónimo de un sentido $k_2$ en WordNet, quitamos todos los sentidos diferentes tanto del concepto $k_1$ como del concepto $k_2$ y en la celda de cruce se indica esta relación.
   d. Si un sentido de $k_1$ es opuesto al sentido $k_2$ en WordNet, quitamos todos los sentidos diferentes tanto del concepto $k_1$ como del concepto $k_2$ y en la celda de cruce se indica dicha relación.

El resultado final de la matriz de emparejamiento entre dos nodos puede ser o un conjunto vacío o un tipo de las posibles relaciones (oposición, sinonimia, hiperonimia, hiponimia) (ver Figura 5). El resultado de la matriz es equivalente a la relación que se evidencia entre el conjunto de documentos clasificados bajo el nodo, con dicha significación contextual, en la JC*f* y el conjunto de documentos clasificados bajo el nodo, con su correspondiente significación contextual, en la JC*d*. La mejor relación teórica entre ambos conjuntos de documentos es la expresada con respecto al orden parcial que indica tanto la «≡» (sinonimia) como la «⊥» (oposición): son mejor que la «⊆» (hiponimia) y la «⊇» (hiperonimia) (BOUQUET 2003, MAGNINI 2002a y 2002b).




*El subdiagrama de una posible significación del nodo etiquetado con «Parto» en la Jerarquía Conceptual Fuente*

| | | |
|---|---|---|
| Obstetricia#1<br>Obstetricia#2 | Parto#1 | Distocia#1 |

*El subdiagrama de una posible significación del nodo etiquetado con «Eutócico» en la Jerarquía Conceptual Diana*

| | | | |
|---|---|---|---|
| Pérdida#1<br>Pérdida#2<br>Pérdida#3<br>Pérdida#4 | Salud#1<br>Salud#2 | Urgencias#i | Eutócico#i |

*Una de las posibles matrices resultantes del emparejamiento de los conceptos «parto» y «eutócico»*

| | Obstetricia#1 | Parto#1 | _Distocia#1 |
|---|---|---|---|
| Pérdida#4 | | | |
| Salud#2 | ⊆ | | |
| Urgencias#i | | | |
| Eutócico#i | | ⊇ | ⊥ |

≡ la relación entre el concepto fuente y el concepto diana es de sinonimia.
⊆ la relación entre el conceptos fuente y el concepto diana es de hiponimia.
⊇ la relación entre el conceptos fuente y el concepto diana es de hiperonimia
⊥ la relación entre el conceptos fuente y el concepto diana es de oposición

**FIGURA 5.** Ejemplo de una posible matriz de emparejamiento de los conceptos «parto» y «eutócicos»

## 4 Discusión y valoración de la propuesta

En este trabajo hemos presentado una propuesta para el diseño de un algoritmo de negociación de significados que permita la interoperabilidad semántica entre dos ontologías locales que se caracterizan por una representación heterogénea pero solapada de una parte común del ámbito de la salud. La interconexión entre jerarquías de conceptos ha sido abordada desde una perspectiva en la que no es preciso trabajar con los documentos asociados a los nodos de éstas. La metodología propuesta se basa en la interpretación lingüística de las etiquetas existentes en las jerarquías, lo que garantiza la identidad semántica de cada una de las jerarquías conceptuales.

El procedimiento cuenta con dos etapas. La primera fase se inicia con la entrada de las etiquetas de una jerarquía de conceptos para construir la fórmula lógica de sus nodos. Esta fase se compone de dos pasos fundamentales: i) el análisis lingüístico morfosintáctico que permite asociar a cada nodo la fórmula lógica que codifica la información que es externa al propio contexto de la jerarquía y que, a modo de símil, es equivalente a la propia interpretación humana de los conceptos que incluye un término; ii) la filtración y combinación de los sentidos asociados a los lemas de las etiquetas respecto al contexto o ubicación en la jerarquía conceptual, para establecer la fórmula lógica de interpretación contextual semántica del nodo, basada en su situación vertical y horizontal en la jerarquía, y para evidenciar la información implícita en la propia etiqueta del nodo. La segunda etapa se centra en la interconexión (*mapping*) entre las jerarquías de conceptos. Para ello se implementa una matriz con la fórmula lógica de interpretación contextual semántica de los dos nodos pertenecientes a la jerarquía conceptual diana y a la jerarquía conceptual fuente. De este modo, se pueden establecer interconexiones entre conceptos con diferente nivel de abstracción. El resultado de la matriz es equivalente a la relación que se evidencia entre el conjunto de documentos que poseen la significación contextual del nodo bajo el que han sido clasificados en la jerarquía conceptual «fuente», y el conjunto de documentos que poseen la significación contextual del nodo bajo el que han sido clasificados en la jerarquía conceptual «diana.»

Sin embargo, falta por dilucidar el cruce de emparejamiento de dos clasificaciones jerárquicas completas, donde deberíamos trabajar con una matriz compuesta por el resultado del emparejamiento de todos los nodos de la «JC diana» con todos los nodos de las «JC fuente». Asimismo, faltaría asociar a cada tipo de relación una posible medida cuantitativa o un grado de estimación, ya que la «JC diana» es una generalización de la «JC fuente» o viceversa, siendo el porcentaje más bajo el que conllevaría la generalización más alta y su repercusión en la recuperación de los documentos.

No obstante, dentro del contexto universitario actual, la propuesta de trabajo establecida tiene una utilidad inmediata para el proceso de enseñanza/aprendizaje de, al menos, las/los alumnas/os de enfermería. En el área de salud, los cambios y nuevas evidencias son casi vertiginosos, lo

que motiva la necesidad de un acceso y consulta de diversas fuentes y recursos de información de forma constante. Este comportamiento se estimula durante el período de formación, pero no sólo se debería pensar en su utilidad durante la fase de teoría en las aulas, sino durante su aprendizaje práctico en instituciones asistenciales. En el ámbito clínico, una de las corrientes que cada vez tiene una mayor influencia y aceptación es la denominada «práctica basada en la evidencia». Es una metodología que busca evitar la generalización a partir de la experiencia no sistematizada, propia o ajena, y obtenida con un número limitado de casos que puede dar lugar a errores. La aproximación clásica basada en la experiencia y consulta entre colegas con frecuencia es ineficaz para solucionar los problemas clínicos concretos y se precisa de la revisión de las fuentes de información (artículos, nuevas recomendaciones de organismo internacionales, indicaciones sobre fármacos, etc.) y de los datos existentes (estadísticas internacionales, nacional o autonómicas, casos atendidos con similar situación, pronóstico y evolución, etc.). Aunque uno de los beneficios de disponer de datos electrónicos es la velocidad de búsqueda, nuestra experiencia en estos años nos muestra que existe una gran dificultad para el acceso y consulta (RICHART 2000). Cada recurso, base de datos o sistema de información indexa su información a partir de su propio vocabulario (controlado o libre), lo que determina la clasificación de información y, por tanto, representa su conceptualización. Las/los alumnas/os precisan ser formados en cada uno de los sistemas de recuperación de la información que deben consultar, no tanto en sus condiciones técnicas sino en los mecanismos que les permiten expresar su necesidad de información de acuerdo con la conceptualización de cada uno de los sistemas de información. En definitiva, para obtener una respuesta que satisfaga la necesidad de información es preciso entender e interpretar los esquemas de clasificación; es decir, su conceptualización, lo que se convierte en un obstáculo para su uso efectivo.

Como conclusión indicar que el modelo propuesto de negociación de significados entre ontologías puede permitir al usuario (alumno) localizar la información que precisa en otra clasificación jerárquica (del dominio de la salud), de forma eficaz y eficiente, sin precisar de un entrenamiento previo (respecto a la conceptualización de cada sistema), ya que partiría de la ontología local con la que ha aprendido o se ha familiarización (el programa de la asignatura) para el conocimiento de su dominio de especialización (enfermería materno-infantil).

## 5 TRABAJO EN PROCESO

Como trabajos futuros, y a partir del análisis cualitativo del algoritmo de interconexión semántica que hemos presentado en este trabajo, hemos planificado: i) Diseñar e implementar computacionalmente la ontología local diana «EMI»; ii) Generar la jerarquía semántica de la versión CIE-9-MC, utilizando técnicas de PLN para la extracción de relaciones semánticas; iii) Desarrollar el algoritmo de interpretación de las etiquetas las je-

rarquías conceptuales fuente y diana; iv) Desarrollar el algoritmo de emparejamiento con las jerarquías conceptuales involucradas; y v) Experimentar con el algoritmo propuesto y evaluar los resultados cualitativos y cuantitativos obtenidos.

## Bibliografía citada


(Bouquet 2002) Bouquet, P.; Donà, A.; Serafín, L.; Zanobini, S. «Contextualized local ontologies specification via CTXML» [recurso electrónico]. En: Proceedings of American Association for Artificial Intelligence (AAAI) workshop on meaning negotiation (MeaN-02). (Edmonton, Canada: AAAI, 2002). <http://dit.unitn.it/~bouquet/papers/AAAI2002-MN.pdf>. [Consulta: 20 octubre 2004].

(Bouquet 2003) Bouquet, P.; Magnini, B.; Serafín, S. «A SAT-based algorithm for context matching». En: Blackburn, P.; Ghidini, Ch.; Turner, R.M.; Giunchiglia, F. (Ed.). *Proceedings of 4th international and interdisciplinary conference on modelling and using context (CONTEXT 2003)*. (Stanford, California: Stanford University 2003). Vol. 2680, p. 66-79.

(Chandrasekaran 1999) Chandrasekaran, B.; Josephson, J.R.; Benjamins, V.R. «What are ontologies, and why do we need them?» *IEEE Intelligent Systems.* (Enero-Febrero 1999), p. 20-26.

(de la Cueva 2001) de la Cueva Martín, A.; Aleixandre Benavent, R.; Rodríguez i Gairín, J.M. *Fonts d'informació en ciències de la salud*. València: Universitat de València, 2001.

(Feliu 2002) Feliu, J., Vivaldi, J.; Cabré, M.T. «Ontologies: a review» [recurso electrónico]. *Sèrie Informes, 34.* Universitat Pompeu Fabra: Instituto Universitario de Lingüística Aplicada, 2002. <ftp://ftp.iula.upf.es/pub/publicacions/02inf034.pdf>. [Consulta: 10 Febrero 2004].

(Gangemi 1998) Gangemi, A.; Pisanelli, D.M.; Steve, G. «Ontology integration: experiences with medical terminologies». En: Proceedings of the Conference: Formal Ontology in Information Systems (1998). <http://www.loa-cnr.it/Papers/fois98r.pdf>. [Consulta: 20 Abril 2004].

(Gruber 1993) Gruber, T.R. «Toward principles for the design of ontologies used for knowledge sharing». En: Guarino, N. & Poli, R. (Ed.). *Formal Ontology in Conceptual Analysis and Knowledge Representation*. Kluwer Academic, in preparation. Original paper presented at the International Workshop on Formal Ontology (1993). Available as Stanford Knowledge Systems Laboratory Report KSL-93-04. <http://ksl-web.stanford.edu/knowledge-sharing/papers/onto-design.rtf>. [Consulta: 10 Febrero 2004].

(Itami 1987) Itami, H. *Mobilizing invisible assets*. Cambridge: Harvard University Press, 1987.

(Lozano 2002) Lozano Tello, A. *Métrica de idoneidad de ontologías*. Tesis Doctoral. Universidad de Extremadura: Departamento de Informática, 2002.

(Magnini 2000) Magnini, B.; Cavaglia,G. «Integrating Subject Field Codes into WordNet». En: Gavrilidou, M.; Crayannis, G.; Markantonatu, S.; Piperidis, S.; Stainhaouer, G. (Ed). *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*. (Athens, 2000), p. 1413-1418.

(Magnini 2002a) Magnini, B.; Serafíni, L.; Speranza, M. «Linguistic based matching of local ontologies». En: Working Notes of the American Association for Artificial Intelligence. Workshop on Meaning Negotiation. (Edmonton, Canadá: AAAI, 2002).

(MAGNINI 2002B) MAGNINI, B.; SERAFÍNI, L.; SPERANZA, M. «Using NLP techniques for meaning negotiation». En: PROCEEDINGS OF VIII CONVENGO AI*IA. (Siena, Italia, 2002).

(MAGNINI 2003) MAGNINI, B.; SERAFÍNI, L.; SPERANZA, M. «Making explicit the semantics». En: PROCECDINGS OF THE WORKSHOP ON HUMAN LANGUAGE TECHNOLOGY FOR THE SEMANTIC WEB AND WEB SERVICES AT ISWC. (Sanibel Island, USA, 2003).

(NONAKA 1994) NONAKA, I. «A dynamic theory of organizational knowledge creation». *Organization Science*, 5 (1) (1994), p. 14-37.

(RICHART 2000) RICHART MARTÍNEZ, M.; CABRERO GARCÍA, J.; TOSAL HERRERO, B.; ROMÁ-FERRI, M.T.; VIZCAYA MORENO, M.F. *Búsqueda bibliográfica en enfermería y otras ciencias de la salud. Bases de datos en Internet*. Alicante: Universidad de Alicante, 2000.

(RODRÍGUEZ 2000) RODRÍGUEZ, H. «Adquisición automática y uso de taxonomías de amplia cobertura». En: WORKSHOP ON AUTOMATIC ACQUISITION OF LINGUISTIC KNOWLEDGE: *tutorial*. (San Millán, 2000). <http://www.lsi.upc.es/~horacio/varios/sanmillan2000.zip> [Consulta: 20 Octuber 2003].

(ROMÁ-FERRI 2004) ROMÁ-FERRI, M.T. *Estudio de interconexión de ontologías del dominio de la salud basado en técnica de procesamiento del lenguaje natural.* Memoria de Suficiencia Investigadora. Programa de Doctorado Aplicaciones de la Informática. Universidad de Alicante: Departamento de Lenguajes y Sistemas Informáticos, 2004.

(SWARTOUT, 1999) SWARTOUT, W.; TATE, A. «Ontologies». *IEEE Intelligent Systems.* (1999, Enero-Febrero), p. 18-19.

(VÁZQUEZ 2004)VÁZQUEZ PÉREZ, S. *Dominios relevantes y su aplicación a la resolución de la ambigüedad léxica*. Memoria de Suficiencia Investigadora. Programa de Doctorado Aplicaciones de la Informática. Universidad de Alicante: Departamento de Lenguajes y Sistemas Informáticos, 2004.